Part-of-Speech Tag Disambiguation by Cross-Linguistic Majority Vote

Aepli, Noëmi ; von Waldenfels, Ruprecht ; Samardžić, Tanja

Part-of-Speech Tag Disambiguation by Cross-Linguistic Majority Vote

Noëmi Aepli[*]
URPP Language and Space
University of Zurich

Ruprecht von Waldenfels[†]
Institute of Computer Science
Polish Academy of Sciences

Tanja Samardžić[*]
URPP Language and Space
University of Zurich

Abstract

In this paper, we present an approach to developing resources for a low-resource language, taking advantage of the fact that it is closely related to languages with more resources. In particular, we test our approach on Macedonian, which lacks tools for natural language processing as well as data in order to build such tools. We improve the Macedonian training set for supervised part-of-speech tagging by transferring available manual annotations from a number of similar languages. Our approach is based on multilingual parallel corpora, automatic word alignment, and a set of rules (majority vote). The performance of a tagger trained on the improved data set of 88% accuracy is significantly better than the baseline of 76%. It can serve as a stepping stone for further improvement of resources for Macedonian. The proposed approach is entirely automatic and it can be easily adapted to other language in similar circumstances.

1 Introduction

Developing natural language processing tools for various languages proves to be of great interest for both, practical applications and linguistic research. Speakers of various languages and varieties increasingly use social media to interact in their own varieties. To make use of these interactions as a relatively easily accessible source of data, we need to be able to process different varieties automatically. However, a great majority of languages of the world lack resources for natural language processing.

With a relatively small number of speakers and weak research infrastructure, Macedonian is one of the languages lacking basic tools for natural language processing. On the other hand, this language is in a convenient position in the sense that it is very similar to other Slavic languages for which more resources are available. We can take advantage of this fact to automatise and facilitate creation of linguistic resources necessary for building tools for automatic processing of Macedonian.

In this paper, we build a part-of-speech tagger for Macedonian. Part-of-speech tagging is a crucial component in a natural language processing pipeline and it is a logical starting point in developing resources for a new language. To obtain a good performance on this task, one needs a sufficiently large corpus with manually annotated tags which can then be used to train a tagger. This is exactly the kind of resource which is often missing (or not easily available) because its development is long, costly and language specific. The current state of language technology allows us to automatise this process to a large degree.

We improve a training set for Macedonian part-of-speech tagging by automatic projection of manual annotation available in other languages. The basis of our method is automatic word alignment, which is widely used in applications for machine translation.

Automatic word alignment has already been used for improving language resources and tools for part-of-speech tagging in the context of supervised (Yarowsky et al., 2001) and unsupervised (Snyder et al., 2008) learning. The success of these techniques strongly depends on the amount of available parallel

[*]{noemi.aepli|tanja.samardzic}@uzh.ch
[†]ruprecht.waldenfels@issl.unibe.ch

Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects, pages 76–84,
Dublin, Ireland, August 23 2014.

corpora for training models for both word alignment and part-of-speech tagging. It is also strongly influenced by the limitations of automatic word alignment which often produces alignment errors, even if it is trained on a large parallel corpus. Our approach to obtaining robust word alignment in a small corpus available for Macedonian is to use a multiple parallel corpus of similar languages. Lexical similarity between the languages is expected to make word alignment easier than for unrelated languages. Combining the information from different languages is expected to cancel out wrong alignments.

2 The Challenge of Developing Resources for Macedonian

Macedonian is an Indo-European language of the Slavic branch. It has around 1.7 million speakers.[1] It is one of the youngest Slavic standard languages, with most of its codification done after the formal declaration of Macedonian as the official language of the Yugoslav Republic of Macedonia in 1944 (Friedman, 2001). Its closest relative is Bulgarian, with whose dialects the Macedonian dialects form a continuum.

2.1 Linguistic Properties

Macedonian belongs to the "Balkan Sprachbund", a famous group of Balkan languages consisting of three Slavic languages (Bulgarian, Macedonian, and some dialects of Serbian), one Romance language (Romanian) and two Indo-European isolates (Greek, Albanian). The members of this group share important structural features developed as a result of areal linguistic contact. The "Sprachbund" features can distinguish the languages belonging to the group from the other languages of the same genetical branch. For example, the Slavic languages belonging to the group differ from all the other Slavic languages in that they do not distinguish cases. To express grammatical relations expressed by case in other Slavic languages, Macedonian and Bulgarian use prepositions (Tomić, 2006). This is an important property in the context of our project because it influences the choice of the direction of automatic word alignment across languages, as it will be shortly described in section 3.2. This property also influences our decision to include in our data set English as the only non-Slavic language (as described in section 2.3).

2.2 Sparse Resources

As far as we know, there is no publicly available part-of-speech tagger for Macedonian at the moment of writing. There are references to morphological resources developed using the NOOJ environment (Ivanovska-Naskova, 2006; Silberstein, 2003). Also, some work on automatic morphological analysis of Macedonian was done in the context of developing an open-source machine translation system (Rangelov, 2011; Peradin and Tyers, 2012).

Most importantly for the current project, a morphologically annotated Macedonian translation of Orwell's 1984 was made available as part of the MULTEXT-East resources (Erjavec, 2012). The annotation in this corpus, however, is incomplete. The main problem is that tokens are assigned all potential part-of-speech tags without disambiguation. Multiple potential tags are assigned to 44,387 tokens, which makes 39% of the whole corpus. Another important problem is missing annotation. There are 4,810 tokens (around 4%) for which there is no annotation at all. The proportion of 43% tokens which lack the crucial information makes this corpus inadequate for training processing tools. To obtain an adequate training set for Macedonian from this corpus, we add the missing information from other languages available in the MULTEXT-East resources with more complete annotation.

2.3 The Overview of our Approach

We take parallel texts for Macedonian (MK), Bulgarian (BG), Czech (CZ), Slovene (SL), Serbian (SR) and English (EN) from the MULTEXT-East corpus (see section 3.1). We select Bulgarian, Czech, Slovene and Serbian as languages closely related to Macedonian. Since these languages are related, they have similar lexicon, grammar and word order. As a result, it can be expected that many words in a parallel text can be aligned as a one-to-one relation, with less cross-linguistic transformations and reordering than in the case of distant languages. In addition to the Slavic languages we also include

[1] https://www.ethnologue.com/language/mkd, 17.04.2014

English because of the fact that Macedonian differs from other Slavic languages (except Bulgarian) in the use of cases. As mentioned above, Macedonian uses analytic prepositional phrases instead of Slavic cases, which makes it closer to languages such as English in this respect.

For each of the five selected languages, manually disambiguated part-of-speech tags are available as part of the MULTEXT-East resources. Moreover, the annotation in different languages can be automatically aligned since the MULTEXT-East corpus consists of translations of the novel "1984" into different languages. All the texts are manually aligned at the level of sentence. Given the sentence alignment, we automatically align Macedonian with the selected languages. We then use word alignments to transfer automatically the annotation found in the other languages to Macedonian. As a next step, we put together all the tags from all the languages, including the available Macedonian tags. This results in a set of part-of-speech candidates for each Macedonian token. We choose the best candidate by a majority vote: the most frequent tag in the set of candidates is chosen as the correct tag. This step relies on the intuition that tags which end up in the candidate set by mistake will not be frequent because their distribution does not depend on the token for which they are candidates. On the other hand, the tags which are truly related to the token in question should be frequent in the set.

The five languages included in the study are not equally close to Macedonian. In addition to the most related languages (Bulgarian and Serbian), we include the data from other Slavic languages (Czech and Slovene) and English to deal with the noise caused by potentially wrong word alignments. We expect that a correct word alignment is more likely to be found in an increased data set. On the other hand, including more languages is not expected to introduce more noise. If word alignments with other languages are wrong, they are not expected to result in repeated tags in the tag candidate set.

Although the general idea is rather intuitive and straightforward, actual realisation of the plan proved technically not trivial. The main difficulty lies in combining word alignment with the original annotation and in cross-linguistic mapping of the manual annotation.

To evaluate the results of the cross-linguistic disambiguation, we provide manual disambiguation for a small section of the Macedonian corpus, which serves as the gold standard. To evaluate how useful our cross-linguistic tag disambiguation is for automatic tagging, we train a tagger on the automatically disambiguated corpus and test it on the portion for which we have provided the gold standard annotation. In the following section, we describe in more detail the decisions taken at each step of our approach.

3 Materials and Methods

As shortly mentioned before, we work with the corpus of the MULTEXT-East resources (Erjavec, 2012), "Multilingual Text Tools and Corpora for Central and Eastern European Languages". The corpus contains the novel "1984" by George Orwell, annotated with part-of-speech tags and further morphosyntactic specifications. It is a parallel corpus available in Macedonian, Bulgarian, Czech, English, Slovene, Serbian and many more. Furthermore, the parallel texts are manually sentence-aligned. The Macedonian corpus was only added in version 4 in 2010. It consists of 113,158 tokens corresponding to 6,790 sentences.

3.1 Multilingual Morphosyntactic Specifications

Morphosyntactic specifications are assigned manually to each token in the corpus. They are similar and largely equivalent across the languages included in the resource, but they are not fully consistent.

Each morphosyntactic definition specifies a value for a number of categories. Each definition consists of a string of characters, where each character specifies the value for one category. These strings can be rather long for words for which many categories need to be encoded. For example, the tag *#Vmia2s----------e* specifies a Macedonian verb form with 15 categories: 1) ***V*** for ***V****erb*, 2) ***m****ain* as *type*, 3) ***i****ndicative* as the verb *form*, 4) ***a****orist* as *tense*, 5) ***2****nd person*, 6) ***s****ingular*, and 7) *perfective* (***e***) as *aspect*. In between, there are no specifications (-) for the subcategories 8) *gender*, 9) *voice*, and 10) *negative*, which could be specified in Macedonian, but have no value in this specific case. Furthermore, there are five subcategories which are not specified for Macedonian but only for other languages, they are marked with a dash too.

Detailed descriptions can be found on the web page of the MULTEXT-East resources.[2]

We notice that the cross-linguistic mapping of the morphosyntactic definitions is more straightforward towards the left-hand side of the definition than towards the right-hand side. For our purpose we only consider the first two letters: the main category and its type (in this example *Vm*). We ignore the information concerning the grammatical categories and reduce the morphosyntactic definitions to relatively coarse part-of-speech tags.

There are 14 main categories (e.g. noun, verb, etc.). Each of these categories can be further specified for the type, but not necessarily. All the combinations of the first two letters in the corpus give a tag set which consists of 58 tags.

Even though morphosyntactic definitions are more consistent across languages for the first two than for the subsequent characters, some variation is found in our tags too. The variations in the subcategories are due to differences in the languages as well as different annotation strategies.

Table 1 shows the categories with the corresponding subcategory *type* across the languages we use. The first and second column of table 1 specify the PoS category to which the types for the six languages are specified. The possible values for the type of the category in one language are separated by a slash (/). The dash (-) means that the type is not specified for that language. A missing entry shows that the whole category is not specified for the language. We can see, for example, that there are three kinds of adjectives in Macedonian: *Af*, *As*, and *Ao*. There are no types in Bulgarian, while the types in other languages overlap with Macedonian only partially. The types which are found in other languages, but not in Macedonian (e.g. *Ag* and *Ap* in Slovenian) cannot be transferred to Macedonian.

		MK	**BG**	**CS**	**SL**	**SR**	**EN**
N	Noun	c/p	c/p	c/p	c/p	c/p	c/p
V	Verb	m/a/o	m/a	m/a/o/c	m/a	m/a/o/c	m/a/o/b
A	Adjective	f/s/o	-	f/s	g/s/p	f/s/o	f
P	Pronoun	p/d/i/s/q r/x/z/g	p/d/i/s/q r/x/z/g	p/d/i/s q/r/x	p/s/d/r/x g/q/i/z	p/d/i/s/q r/x/z/g	p/s/q/r x/g/t
R	Adverb	g/a/v	g/a	g	g/r	g/z/a/v	m/s
S	Adposition	p	p	p	-	p	p/t
C	Conjunction	c/s	c/s	c/s	c/s	c/s	c/s
M	Number	c/o/l/s	c/o	c/o/m/s	c/o/p/s	c/o/m/l/s	c/o
I	Interjection	-	-	-	-	-	-
Y	Abbreviation	-	-	n/r	-	n/r	-
X	Residual	-	-	-	f/t/p	-	-
Q	Particle	s/c	z/g/c/v/q/o	z/q/o/r	-	c/a/o/r	
D	Determiner						d/i/s/g
T	Article						

Table 1: Cross-linguistic mapping of part-of-speech tags in our data set.

3.2 Automatic Word Alignment

The MULTEXT-East corpus contains manual sentence alignment for each language pair. We extract the information about sentence alignment between Macedonian and the five languages included in our study.

Given the sentence alignment, we word align each of the parallel texts using GIZA++ (Och and Ney, 2003). As it is required by the input format for GIZA++, we remove sentence boundaries in the cases where sentence alignment is not one-to-one. For example, if two English sentences are aligned with one Macedonian sentence, we remove the boundary between the two English sentences. We then restore the sentence boundaries in the alignment output so that we can identify the sentences in the original annotated corpus and retrieve the annotation.

For each pair of languages, word alignment can be performed in two directions. One language is considered as the source and the other as target. The choice of the alignment direction can have an important influence on the resulting alignment (Och and Ney, 2003; Samardžić and Merlo, 2010). The influence of the alignment direction on the results follows from the formal definition of word alignment

[2]http://nl.ijs.si/ME/, 24.06.2014

in the practical implementation. Since alignment is a single-valued function which assigns to each target language word exactly one source language word, many-to-one alignments are only possible in one direction: multiple target language words can be aligned with one source language word, but not the other way around.

The performance of the programs for automatic word alignment is not perfect. To obtain more reliable alignment, researchers usually take the intersection of both directions as the resulting alignment. This technique yields very reliable alignments reaching a precision of 98.6%. However, since it allows only one-to-one alignment, it necessarily leaves a good proportion of words unaligned (recall as low as 52.9%) (Padó, 2007).

Since our corpus is small, we need to obtain as many word alignments as possible. Thus we do not use the intersection of both alignments, but we use the full output of one alignment direction. It follows from the formal definition of alignment that all target words need to be aligned, which necessarily increases the recall, but potentially at the cost of precision.

To obtain a better precision, we choose the more suitable direction of alignment. Since the many-to-one mappings are possible only from the target language to the source language, we choose the alignment direction for each pair of languages so that the target language is the more analytic one. In all Slavic pairs, Macedonian is the target, due to the fact that it uses analytic prepositional expressions where other Slavic languages use single words in a particular case. In the pair English-Macedonian, the target language is English, because its forms are more analytic than in Macedonian.

3.3 Combining Information from All Languages

Given the word alignment, we replace each word of the other languages (OL) which is aligned to a Macedonian word with its corresponding part-of-speech tag retrieved from the original manually annotated corpus. Table 2 illustrates the resulting data structure. The first column in the table is the sentence ID, the second the Macedonian word. In the next columns the part-of-speech information is stored: first the Macedonian tags and then the tags projected from other languages. Language code is given before "#" and the full morphosyntactic definition found in the language in question after "#".

As it can be seen in Table 2, none, one, or several tags can be specified for each language. In the first example, there is exactly one tag for every language. In the second example, the part-of-speech information in English is missing because there was no alignment between the Macedonian word "со" and any English word. This is the case for all five other languages in the last example, where the tags are specified only for Macedonian.[3] The third example shows the opposite, with one PoS tag for each other language, but none for Macedonian.

ID	Word	MK PoS	OL PoS
1.1.1.1	јасен 'clear'	mk#Af	bg#AM cs#Af en#Af sl#Ag sr#Af
1.1.1.2	со 'with'	mk#Sp	bg#SP cs#Rg en sl#Si sr#Sp
1.1.1.2	Винстон 'Winston'		bg#Np cs#Np en#Np sl#Np sr#Np
2.7.2.3	едно 'one'	mk#C- mk#Rg mk#Mc	bg#VM cs#Mc en#Di sl#Ap sr#Vm
1.1.11.2	што 'what'	mk#Pq mk#Pr mk#C- mk#Q- mk#Rg mk#I	bg cs en sl sr

Table 2: Macedonian text with PoS tags of aligned words of other languages

3.4 Choosing the Best Candidate

Having collected sets of possible tags for each Macedonian word, the next step is to choose the best tag.

The general idea is to take into consideration all the tags of all languages that are given for one word and choose the most frequent of them as the correct tag for Macedonian. As the tags do not match

[3]Note that alignments are not missing in the technical sense in the case of Slavic languages. According to the formal definition of alignment discussed above, all Macedonian words need to be aligned in the direction that we chose. The fact that there is no alignment in our data means that the Macedonian word is aligned with the special "NULL" word in other Slavic languages in this case. This special word is added to each sentence of each source language in the process of alignment, so that the target language words for which there are no corresponding words in the source language can be aligned too.

completely (see section 3.1), the chosen tag has to be checked for validity. In other words, we check if the most frequent tag is a valid tag for Macedonian according to the MULTEXT-East specifications.

For the task of choosing the best tag, we define a set of if-then rules. We apply an outer structure of three if/else statements checking how many tags are given for Macedonian: one, zero or several. If exactly one tag is given, we choose it as the best candidate. The latter two cases include further checks taking into account the number of specified tags of the other languages (zero or several) as well as the number of most frequent tags (the maximum). The former check is necessary because of the cases where there are zero tags in Macedonian. If there are no tags in other languages either, we have to assign a "dummy tag". The dummy tag is the most frequently occurring tag in the original annotation for Macedonian. This is the *Nc* (common noun) tag in our case. The latter check, the number of maxima, is done because more than one tag could have the same frequency. In cases where the competition between the tags remains unresolved because of no matchings and/or sparse data, we reduce the tag to make it less specific. We ignore the type, that is, the second letter of the tag, which leaves us with only the category. Even this approach does not solve all the decision problems. If this is the case we have two procedures: if there is no tag information coming from any language, we assign a "dummy tag". In the second case, where we can not decide but we do have some information in Macedonian, we randomly choose one of the given Macedonian tags. The cases in which we had to apply some additional heuristics (comparing reduced tags, random choice and dummy tag) because there was not one single most frequent tag constitute around 10%. The decision process for choosing the best candidate is given in more detail in the pseudocode "Algorithm 1".

Consider, for example, the fourth entry in Table 2, "едно". There are three tags for Macedonian, which means it satisfies the third condition of the outer if/else structure (more than 1 MK PoS tag). Next, the most frequent tag considering all the given PoS tags of all the languages is searched. As described in Section 3.1, we only take into account the first two letters (category and type) of a given morphosyntactic definition. In this case, we have the following tags with the corresponding frequencies: (MC : 2), (VM : 2), (C : 1), (AP : 1), (DI : 1), (RG : 1). Looking for the maximum, we find two tags with the same frequency (2): MC and VM. Because there is more than one maximum, we check for each of the two tags if they are identical to one of the Macedonian tags. In this case, the test is true for MC (cardinal numeral). This is one of the maxima **and** one of the Macedonian tags, therefore the winner.

3.5 Training a Tagger

To asses whether disambiguating part-of-speech tags as described in the previous sections is useful for training a statistical part-of-speech tagger, we divide our data set into a training and test portion. We train a tagger on the training portion of the disambiguated corpus and we measure its performance on the test set. We use the BTagger (Gesmundo and Samardzic, 2012), since it has good generalisation capacities, which makes it suitable for small data sets. Furthermore, it does not need any manually constructed morphological dictionaries and it can be used for any language.

4 Evaluation

To evaluate both our disambiguation method and the performance of the tagger on the disambiguated corpus, we chose an arbitrary sample section of the corpus as the test set. The sample included 9,954 tokens (around 10% of the whole corpus), out of which 616 were missing annotation, and 3,231 were not disambiguated. We manually add the missing tags and disambiguate the ambiguous ones. In this way, we obtain the gold standard for the evaluation.

4.1 The baseline

We compare both, the success of our cross-linguistic disambiguation and the performance of the tagger with a baseline. To define the baseline, we use a simple heuristic which allows us to disambiguate Macedonian tags without cross-linguistic information: we take the first tag in the list as the correct one. In the case of missing tags, we add NC (common noun), which is the most frequent tag in the corpus. We run the tagger on the corpus disambiguated in this way, which gives us the baseline performance.

Algorithm 1 Find the best PoS-tag for an MK word given MK, BG, CS, EN, SL and SR tags

```
1: if number of MK-PoS-tags = 1 then
2:     result ← this MK-PoS-tag
3: else if number of MK-PoS-tags = 0 then
4:
5:     if number of OL-PoS-tags = 0 then
6:         result ← dummy-tag
7:     else if number of OL-PoS-tags > 0 then
8:
9:         if 1 maximum then
10:            result ← maximum (→ to be checked whether it is a valid MK-tag)
11:        else if >1 maximum then
12:            result ← dummy-tag
13:        end if
14:    end if
15: else if number of MK-PoS-tags > 1 then
16:
17:    if 1 maximum then
18:
19:        if maximum = one of MK-PoS-tags then
20:            result ← maximum
21:        else if reduced PoS-tag = one of MK-PoS-tags then
22:            result ← MK-PoS-tag with the same category like the maximum
23:        else if maximum not in MK-PoS-tags then
24:            result ← random choice of available MK-PoS-tags
25:        end if
26:    else if > 1 maximum then
27:
28:        for candidate in maxima do
29:
30:            if candidate = one of MK-PoS-tags then
31:                result ← candidate
32:            else if candidate not one of MK-PoS-tags then
33:                reduce candidate to 1 letter
34:                if reduced candidate = one of reduced MK-PoS-tags then
35:                    result ← not-reduced MK-PoS-tags
36:                else
37:                    result ← random choice of available MK-PoS-tags
38:                end if
39:            end if
40:        end for
41:    else if number of OL-PoS-tags = 0 then
42:        result ← random choice of available MK-PoS-tags
43:    end if
44: end if
```

4.2 Results and Discussion

Table 3 shows the accuracy of cross-linguistic disambiguation and tagging in comparison with the baseline. The second column shows the agreement between manual disambiguation (the gold standard) and automatic disambiguation in the two settings.

We can see that our simple heuristics alone provide some correct disambiguation. Roughly half of the 43% of tags which are potentially wrong in the original corpus (because they are not disambiguated or because they miss annotation) are correctly disambiguated by the baseline heuristics. This gives the baseline disambiguation accuracy of 78%. Adding the information from other languages improves the accuracy of automatic disambiguation to 87%.

Accuracy (%)	Disambiguation	BTagger	
Baseline	78	All	77
		Known	76
		Unknown	77
Cross-linguistic Majority Vote	87	All	88
		Known	88
		Unknown	91

Table 3: The accuracy of disambiguation and tagging compared with the gold standard.

When trained on the corpus disambiguated in the baseline setting, the tagger's accuracy is 77%, while its accuracy is improved to 88% when it is trained on the corpus disambiguated using our cross-linguistic majority vote.

It is important to note that the tagger's performance improves more than the disambiguation accuracy compared to the baseline (77% to 88% vs. 78% to 87%). The tagger outperforms the direct disambiguation in the cross-linguistic setting. This means that eliminating wrong tags from the training set allows the tagger not only to learn better correct tags, but also to come up with generalisations and provide a more robust output. Although it assigns learned wrong tags to the words seen in the training set (accuracy on known words 88%), it uses the learned generalisations to predict more correct tags on the words unseen in the training set (accuracy on unknown words 91%).

5 Conclusion

We have presented a method for improving resources in a new language using the existing resources in similar languages and state-of-the art language technology. We evaluated our method as applied to Macedonian, a low-resource Slavic language, closely related to other Slavic languages with more available resources.

By cross-linguistic annotation projection, we improved the existing annotation, assigning the correct tag to two thirds of potentially wrong part-of-speech tags in the original corpus. The performance of a tagger trained on the disambiguated corpus reaches 88% accuracy. This is not a satisfying performance in itself, but this tagger is the first trained and evaluated tool for Macedonian. Another important outcome of our experiments is the fact that an improved training set allows a tagger to develop crucial generalisations and to provide a more robust output. This finding can be useful for further improvement of the resources not only in Macedonian, but in other low-resource languages too.

The presented approach to improving annotated language resources across languages is entirely automatic. It can be applied to any other language in similar circumstances. Instead of repeating the same kind of costly, time-consuming manual work in each new language, our approach makes use of available annotations by transferring them automatically from one language to another.

Acknowledgements

The work presented in this paper is supported by the URPP Language and Space, University of Zurich and the Swiss National Science Foundation. Training data annotation was co-financed by the Slavic Institute of Bern University. Many thanks to Andrea Gesmundo for valuable comments and suggestions.

References

Tomaž Erjavec. 2012. MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages. In *Language Resources and Evaluation*, volume 46, pages 131–142.

Victor A. Friedman, 2001. *Facts About The World's Languages: An Encyclopedia of the World's Major Languages, Past and Present*, chapter Macedonian, pages 435 – 439. The H. W. Wilson Company New York and Dublin.

Andrea Gesmundo and Tanja Samardzic. 2012. Lemmatisation as a tagging task. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 368–372, Jeju Island, Korea, July. Association for Computational Linguistics.

Ruska Ivanovska-Naskova. 2006. Development of the First LRs for Macedonian: Current Projects. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 1837–1841. European Language Resources Association (ELRA).

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. In *Computational Linguistics*, volume 29, pages 19–51.

Sebastian Padó. 2007. *Cross-Lingual Annotation Projection Models for Role-Semantic Information*. Ph.D. thesis, Saarland University.

Hrvoje Peradin and Francis Tyers. 2012. A rule-based machine translation system from Serbo-Croatian to Macedonian. In *Proceedings of the Workshop Free/Open-Source Rule-Based Machine Translation*, pages 55 – 62, Gothenburg, Sweden.

Tihomir Rangelov. 2011. Rule-based machine translation between Bulgarian and Macedonian. Universitat Oberta de Catalunya.

Tanja Samardžić and Paola Merlo. 2010. Cross-lingual variation of light verb constructions: Using parallel corpora and automatic alignment for linguistic research. In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground*, pages 52–60, Uppsala, Sweden. Association for Computational Linguistics.

Max Silberstein. 2003. NooJ Manual. Available at www.nooj4nlp.net.

Benjamin Snyder, Tahira Naseem, Jacob Eisenstein, and Regina Barzilay. 2008. Unsupervised Multilingual Learning for POS Tagging. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1041–1050, Honolulu. Association for Computational Linguistics.

Olga Mišeska Tomić. 2006. *Balkan Sprachbund Morpho-syntactic Features*. Springer, Dordrecht, The Netherlands.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the 1st international conference Human Language Technology*, pages 161–168, San Diego, CA. Association for Computational Linguistics.